DOCUMENT RESUME

ED 064 356 TM 001 524

AUTHOR Kleinke, David J.
TITLE The Accuracy of Estimated Total Test Statistics. Final Report.
SPONS AGENCY National Center for Educational Research and Development (DHEW/OE), Washington, D.C.
BUREAU NO BR-I-B-070
PUB DATE Mar 72
GRANT OEG-2-710070
NOTE 11p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Academic Achievement; *Item Sampling; Predictive Validity; Research Methodology; *Sampling; *Test Construction; *Test Results

ABSTRACT

In a post-mortem study of item sampling, 1,050 examinees were divided into ten groups 50 times. Each time, their papers were scored on four different sets of item samples from a 150-item test of academic aptitude. These samples were selected using (a) unstratified random sampling and stratification on (b) content, (c) difficulty, and (d) both. There were no systematic relationships between method of sampling and accuracy or stability (defined in terms of the means and variances of the distributions of the estimates) of estimated total-test mean or variance. Implications for both generalizability theory and item sampling methodology are discussed. (Author)

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ABSTRACT

In a post-mortem study of item sampling, 1,050 examinees were divided into ten groups fifty times. Each time, their papers were scored on four different sets of item samples from a 150-item test of academic aptitude. These samples were selected using (a) unstratified random sampling and stratification on (b) content, (c) difficulty, and (d) both. There were no systematic relationships between method of sampling and accuracy or stability (defined in terms of the means and variances of the distributions of the estimates) of estimated total-test mean or variance. Implications for both generalizability theory and item sampling methodology are discussed.

ED 064356

Final Report

Project No. IB070
Grant No. OEG-2-710070

David J. Kleinke
Syracuse University
337 Huntington Hall
Syracuse, New York 13210

THE ACCURACY OF ESTIMATED TOTAL TEST STATISTICS

March, 1972

TM 001 524

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

Office of Education

National Center for Education Research and Development

Final Report

Project No. IB070
Grant No. OEG2-710070

THE ACCURACY OF ESTIMATED TOTAL TEST STATISTICS

David J. Kleinke
Syracuse University
Syracuse, New York

The research reported herein was performed pursuant to a grant with the U. S. Office of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Point of view or opinions do not, therefore, necessarily represent official Office of Education position or policy.

U. S. DEPARTMENT OF

HEALTH, EDUCATION, AND WELFARE

Office of Education

National Center for Educational Research and Development

Introduction

Increasingly, psychometricians have been suggesting that matrix sampling be employed in the norming of tests. "Matrix sampling" is that technique under which an examinee responds to only a sample of the items that comprise the full test. The sampling is across both items and examinees, hence "matrix sampling." The term"item sampling," which also has been used to denote the technique, is here defined more narrowly in that it refers to the sampling of only the items. In any event, the results obtained from examinee samples' responding to the item samples are used to generate estimates of the performance that would have been obtained if all of the examinees had been presented with all of the items.

To date, little attention has been paid to the manner in which the items are selected for allocation to item sample, that is, which items are grouped together. In nearly all previous studies of matrix sampling, the item sampling has been done at random without replacement, employing unstratfied selection methods. The rationale underlying generalizability theory, however, suggests that some form of stratified-random sampling would yield better estimates of total-test statistics than would unstratified random sampling. "Stratified-random" sampling is here defined as that in which (a) items are grouped on some a priori basis, such as their content or predicted difficulty, and (b) the item samples are selected in such a manner that they will contain similar proportions of items of like content, difficulty, or both. This study, then, was directed toward the question, "Are estimates of total-test mean and variance arising from stratified-random item sampling more accurate or stable than estimates based on simple (unstratified) random sampling?"

Related Research

Cronbach and his associates (Cronbach, Schönemann, and McKie, 1965; Rajaratnam, Cronbach, and Gleser, 1965) have presented an argument that generalizability from one test to another is increased when the tests' items are selected on a stratified-random basis. Strictly, "generalizability" is the relationship between scores on a test and scores on one of a family of similar tests. The present study investigated empirically, with the actual test score data, whether or not generalizability from a sample of items to the finite number of items comprising the total test is not also enhanced by selecting the items for part-test membership on a stratified-random basis. This relationship between generalizability theory and matrix sampling has been previously pointed out by Osburn (1967; 1968). Cronbach et al (1965) used hypothetical data to test the theory. Their examinations were composed of items that varied in difficulty and, through manipulation of the magnitudes of inter-item correlations, had the appearance of differing in content. They found that generalizability, as approximated through internal-consistency procedures, was least for the unstratified tests, next largest for those stratified on the basis of item difficulty, still larger for the test stratified on the basis of content, and greatest for those stratified on the basis of both

content and difficulty.

The first full-fledged *post hoc* empirical study of matrix sampling was that of Lord (1962). He selected ten seven-item part-tests from among the first seventy items of a vocabulary test taken by 1,000 college seniors. The sampling was from these items in order to avoid problems of interpretation that might have resulted from including "not reached" items. Item sampling was performed at random, with replacement. Lord commented that sampling without would have been preferable since (a) sampling with replacement resulted in having some items' being selected more than once and others omitted, and (b) sampling without replacement would not have violated any of the underlying mathematical assumptions (p. 262). All subsequent investigators have sampled both items and examinees without replacement. The examinee-sample groups were selected from the norming sample randomly without replacement. The papers were scored for both item-sample (part-test) scores and total-test scores. Then, total-test arithmetic means and variances were estimated, using equations that were based on those of Lord (1960, Table 2). These equations or their algebraic equivalents have been employed by nearly all subsequent investigators. Plumlee (1964) provided the first study in which the Lord methodology was used with non-overlapping item samples. She sampled from among a thirty-item punctuation test and 200 applicants for clerical positions. Cook and Stufflebeam (1967) varied the size of examinee- and item samples with a pool of 1,239 college students who had responded to a 115-item achievement test in health education. In both of these studies as well as in that of Lord (1962), the item-sample-based estimates of the total-test mean and variance were generally closer to the actual total-test mean and variance than were estimates derived from having samples of fewer examinees scored for the total test. These later estimates were used for comparison because they were based on as many responses as were those obtained through matrix sampling.

The content areas of the tests employed by Lord (vocabulary) and Plumlee (punctuation) were probably relatively homogeneous, but the college-level achievement tests used by Cook and Stufflebeam could probably have been divided into part-tests on the basis of content area. Moreover, even if there were a test that tapped just one mental trait, its items would probably differ in difficulty. This suggests that item samples be selected not on a simple, unstratified, random basis, but rather that stratification on the basis of both content and difficulty be used to make each item sample resemble as closely as possible the total-test for which estimates are sought.

Kleinke (1969) attempted to investigate the relative accuracy of estimates based on item-samples formed in four different ways. His test was a one hundred-item test of verbal ability that had been administered to more than 150,000 twelfth-grade pupils with college aspirations. Forty of the items were verbal analogies (VAs), and thirty each were sentence completions (SCs) and same-opposites (SOs). Additionally, pretests produced *a priori* estimates of item difficulty. Four different sets of ten ten-item samples were selected. The first set was drawn with simple random sampling. The second was stratified on the basis of estimated item difficulties. Item-type, or content, was the basis for stratification

for the third set of samples, and the fourth was selected after stratification on both difficulty and content. The examinee samples of 105 examinees each were selected randomly without replacement. The hypothesis, that the "random" (R) sample would produce the least accurate estimates of total-test mean and variance, the "difficulty" (D) sample the next most accurate, the "content" (C) sample the next, and the difficulty-by content" (DxC) the most, was not supported. It was argued that the failure to support this hypothesis was due to inadequate design. It was suggested that an arbitrarily large number of different sets of item samples be drawn under each of the four sampling strategies and that the distributions of the estimated statistics derived therefrom be observed.

Objective

It was the purpose of the present study to follow the suggestions of Kleinke (1969) to investigate the effect of type of item sampling on the accuracy and stability of the resulting estimates of total-test mean and variance. A number of sets of item samples were drawn from the same total test, using four different sampling strategies. Under the first strategy, items were sampled on a purely random basis, without replacement. The second strategy was to stratify the item samples on the basis of a priori estimates of item difficulty. Under the third strategy, content (item-type) was the basis for stratification. For the fourth set of samples, stratification was on both difficulty and content.

Following the theory and results of Cronbach et al (1965) and Rajaratnam et al (1965) it was expected that the R samples would yield the poorest estimates of total-test mean and variance; the estimates based on difficulty-stratified item samples would be next best; those based on the C samples, the next best; and the estimates arising from DxC samples would provide the most accurate and stable estimates. "Accuracy" was determined by comparing the mean of the distributions of the estimates with the known population values; "stability," by examining the variances of those distributions.

Method

Sampling Units

The test was the 150-item test of which Kleinke's (1969) one hundred items were the first portion. These were augmented by the inclusion of fifty arithmetic reasoning (AR) questions, which were items 101-150 in the original examination. The problem of including "not reached" items that was encountered by Lord (1962) was not present with these data: no item was omitted by more than two percent of the examinees.

Items were assigned to difficulty strata in the following manner. The frequency distribution of the a priori estimates of item difficulty was divided such that there were 30 items designated "very easy" (VE); 40, "moderately easy" (ME); 50, "moderately difficult" (MD); and 30, "very difficult" (VD). The number in each category were chosen to be proportional to the numbers in the content strata. There were 19 items

whose estimated difficulty indices fell at one of the three cutting points. They were arbitrarily assigned to one of the cells in order to maintain the 3-4-5-3 ratio within content area. Numbers of items by difficulty stratum and content area are presented in Table 1. That there

Table 1. Numbers of items in difficulty strata, by content area

Designation	Estimated Difficulty	Content Area				Total
		SO	VA	SC	AR	
VE	.59-.72	7	10	7	6	30
ME	.51-.59	7	11	7	15	40
MD	.43-.51	11	9	10	20	50
VD	.25-.43	5	10	6	9	30
	Total	30	40	30	50	150

was no apparent relationship between content and distribution of estimated item difficulty was supported with a chi-square goodness-of-fit test ($\chi^2 = 5.84$, d.f. = 9, $.90 > p > .75$).

Items were assigned to item sample using four different sampling strategies. Under the first (R), items were selected at random without replacement. The second strategy, to select the D samples, restricted each item sample to three each VE and VD items, four ME items, and five MD items. In similar fashion the C samples contained three each SC and SO items, four VAs, and five ARs. Item allocation for the DxC samples is presented in Table 2. In all cases item selection was performed at random, without replacement, using an IBM/50 computer.

The examinees were the 1,050 used by Kleinke (1969). To simplify computations, their responses to the 150 items were stored binarally on magnetic tape.

Computations

The following steps were repeated fifty times:

1. Ten nonoverlapping examinee-samples of 105 examinees were drawn, designated 1, 2, ..., 10.

2. Four sets of ten 15-item item samples were selected, designated R1, R, ..., R10, D1, D2, ..., D10, C1, C2, ..., C10, DxC1, DxC2, ..., DxC10.

Table 2. Allocation of items to item-sample for content-by-difficulty samples

Content	Difficulty	Item-sample 1	2	3	4	5	6	7	8	9	10
SO	VE	1	1	1	1	1	1	0	0	1	0
	ME	0	0	0	0	1	1	2	2	0	1
	MD	1	1	1	1	1	1	1	1	2	1
	VD	1	1	1	1	0	0	0	0	0	1
VA	VE	1	1	1	1	1	1	1	1	1	1
	ME	1	1	1	1	1	1	1	1	2	1
	MD	1	1	1	1	1	1	1	1	0	1
	VD	1	1	1	1	1	1	1	1	1	1
SC	VE	1	1	1	1	0	0	1	1	0	1
	ME	1	1	1	1	1	1	0	0	1	0
	MD	1	1	1	1	1	1	1	1	1	1
	VD	0	0	0	0	1	1	1	1	1	1
AR	VE	0	0	0	0	1	1	1	1	1	1
	ME	2	2	2	2	1	1	1	1	1	2
	MD	2	2	2	2	2	2	2	2	2	2
	VD	1	1	1	1	1	1	1	1	1	0

3. Each examinee's responses were "scored" (totaled) for the appropriate four item samples.

4. The forty distributions of item-sample scores were generated and their means and variances computed.

5. Item difficulty indices (proportion correct) were established for each item under each sampling rule. For each item sample, the sum of item variances was computed.

6. For each of the four sampling rules, the estimated total-test mean ($\hat{T}$) and variance ($\hat{s}^2_T$) were computed, using the algebraic equivalents of Lord's (1960, Table 2) equations.

Following these fifty repetitions, there were thus four distributions (R,D,C, and DxC) of $\hat{T}$ and $\hat{s}^2_T$. All 1,050 examinees' responses were also scored for the total test to obtain the criterion mean and variance.

Results

Means and variances of the estimated total-test mean and variance are presented in Table 3. As can be seen, there were no systematic differences

Table 3. Means and variances of estimated total-test statistics, by item-sampling rule

Statistic	Sampling rule			
	Random	Difficulty	Content	D x C
	Estimated mean*			
Mean	75.52	75.38	75.37	75.53
Variance	.23	.28	.23	.25
	Estimated variance**			
Mean	741.14	740.60	739.82	738.78
Variance	894.74	747.60	955.93	752.46

* Criterion mean = 75.43

** Criterion variance = 743.81

in the stability of the estimates (as measured by their variance) or the accuracy of the estimates (as measured by comparing their mean with the criterion value).

Conclusions

It must be concluded that, with the data set at hand, item-sampling strategy made no difference in accuracy or stability of the estimated total-test mean or variance. For the test used and the examinees who responded to it, unstratified random samples yielded estimates that were, in general, no more or less accurate or stable than those that resulted from stratified item samples.

This conclusion was reached after several checks were made on the methodology, in order to rule out some plausible competing hypotheses. For instance, a check was made on the sampling and computational procedures themselves. No anomolies were discovered. That is, the items and these data were numerically correct. Additionally, the scoring of the examinees' responses was found to be accurate.

It could also have been that the stratification on content was effective but that on difficulty was not. To check this, the R- and D-samples and the C- and DxC-samples, respectively, were combined, and then the opposite pairings were made. Shoemaker and Osburn (1968), for instance, suggest that stratification on difficulty is more important than that on content. The means and variances of the grouped estimates were then

examined. Still, no differences were uncovered.

Also, the item difficulty indices themselves were examined. They were found to be stable. That is, the original, a priori, estimates ranked the items' difficulties in virtually the same order as did the item-sampled results. These item difficulties were distributed approximately normally with a mean of about 0.5; more than two-thirds of them were orginally estimated to be between 0.4 and 0.6.

Likewise, interitem correlations were generally within the "moderate positive" range. The total-test and its examinees were, therefore, typical of most well-constructed tests of academic aptitude: items of moderate difficulty with modest, but positive intercorrelations, even after the inclusion of the arithmetic reasoning items in a deliberate attempt to decrease those intercorrelations. That the interitem relationships was high was borne out by an investigation of the Kuder-Richardson Formula 20 reliability estimates for the various item samples. They were essentially the same in central tendency and dispersion for all four sampling rules.

Further research is indicated. This should take two different directions.

The first of these, suggested at the time of the interim report on this project and subsequently also suggested by Shoemaker (personal communication), would be a Monte Carlo study. A number of studies similar to that being reported here should be conducted using generated data whose interitem correlations and item difficulty indices could be manipulated, to uncover which characteristics, if any, of a test affect the accuracy and stability of total-test estimates.

The second line of investigation parallels this. It would be a survey of typical standardized and locally-constructed tests to see just what the distributions of item difficulties and interitem correlations are. The Cronbach et al (1965) and Shoemaker and Osburn (1968) results were based on artificially generated data, with such parameters as rectangularly-distributed item difficulty indices ranging from 0.1 through 0.9. Just how realistic such parameters are is an obvious area for investigation. The point is that stratification may be important, but only under conditions that simply do not exist with actual tests.

Finally, that no differences were observed in the present study can be interpreted to imply that one may stratify if he wishes. If item sampling is to be used for criterion-referenced measurement, or even for the more traditional norming of a test, content-stratification, at least, might lend greater acceptibility to a test. Since the stated or implied purpose of matrix sampling generally is to make the test-taking task more acceptable to the examinees and their administrators, it could be that the increased face validity that could result from stratification would enhance the task's acceptibility.

References

Cook, D. L. and Stufflebeam, D. L. Estimating test norms from variable size item and examinee samples. Journal of Educational Measurment, 1967, 4, 27-33.

Cronbach, L. J., Schönemann, P., and McKie, D. Alpha coefficients for stratified-parallel tests. Educational and Psychological Measurement, 1965, 25, 291-312.

Kleinke, D. J. A linear-prediction approach to developing test norms based on item-sampling. Unpublished doctoral dissertation, State University of New York at Albany; Albany, 1969.

Lord, F. M. Use of true-score theory to predict moments of univariate and bivariate observed-score distributions. Psychometrika, 1960, 25, 325-340.

Lord, F. M. Estimating norms by item-sampling. Educational and Psychological Measurement, 1962, 22, 259-267.

Osburn, H. G. A note on the design of test experiments. Educational and Psychological Measurement, 1967, 27, 797-802.

Osburn, H. G. Item sampling for achievement testing. Educational and Psychological Measurement, 1968, 28, 95-104.

Plumlee, L. B. Estimating means and standard deviations from partial data -- An empirical check on Lord's item sampling technique. Educational and Psychological Measurement, 1964, 24, 573-630.

Rajaratnam, N., Cronbach, L. J., and Gleser, G. Generalizability of stratified parallel tests. Psychometrika, 1965, 30, 39-56.

Shoemaker, D. M. Further results on the standard errors of estimate associated with item-examinee sampling procedures. Journal of Educational Measurement, 1971, 8, 215-220.

Shoemaker, D. M. and Osburn, H. G. An empirical study of generalizability coefficients for unmatched data. British Journal of Mathematical and Statistical Psychology, 1968, 21 (Part 2), 239-246.